Columbia | Engineering
The Fu Foundation School of Engineering and Applied Science

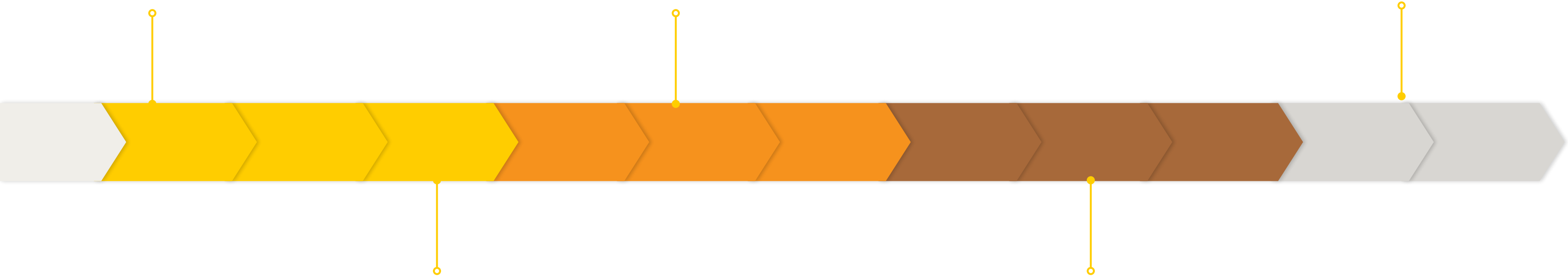
Recommendation Machines for Movie Lovers and Movie Makers

House Targaryen

Sara Ma(ym2841), Hanrui Yu(hy2716), Qiming Feng (qf2155),

Ting Lei(tl3101), Linzi Guan(lg3183), Xuanyu Li (xl3116)

Table of Contents



Idea and Goal

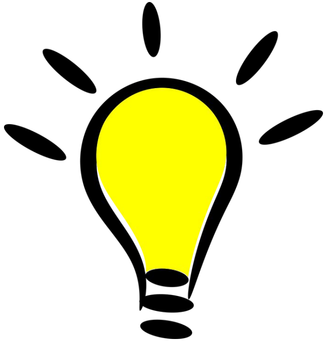
Data Selection and Cleaning

Exploratory Data Analysis & Visualization

Movie Recommendation Machine

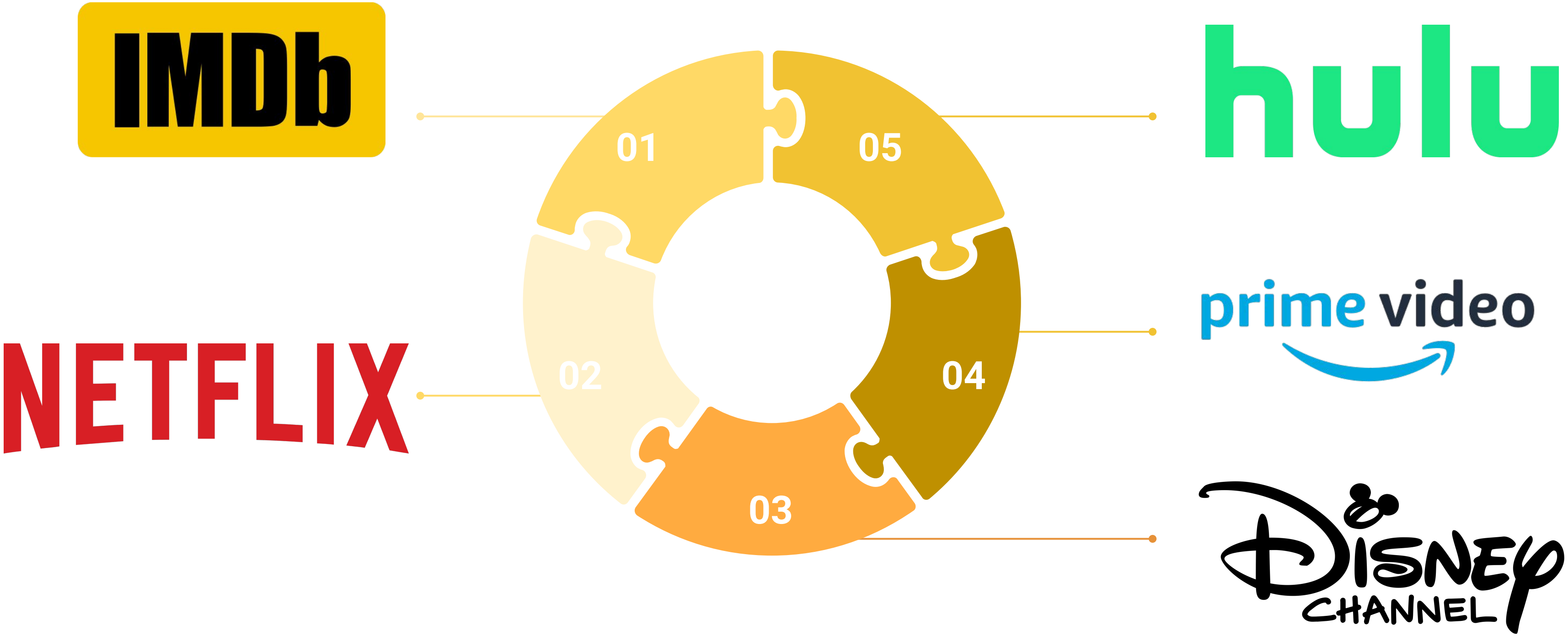
Movie Popularity Prediction

Introduction

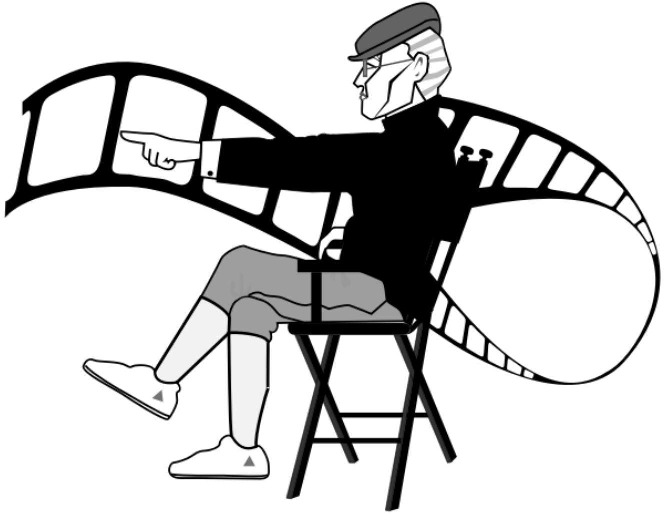


redbox
OVER 200 movies in each box
ICE CUBE KEVIN HART
Ride Along
LOOK ON TOUCH SCREEN FOR MORE TITLES
NEW RELEASE movies & games
call 1-866-REDBOX3
HoBbIT

Data Processing



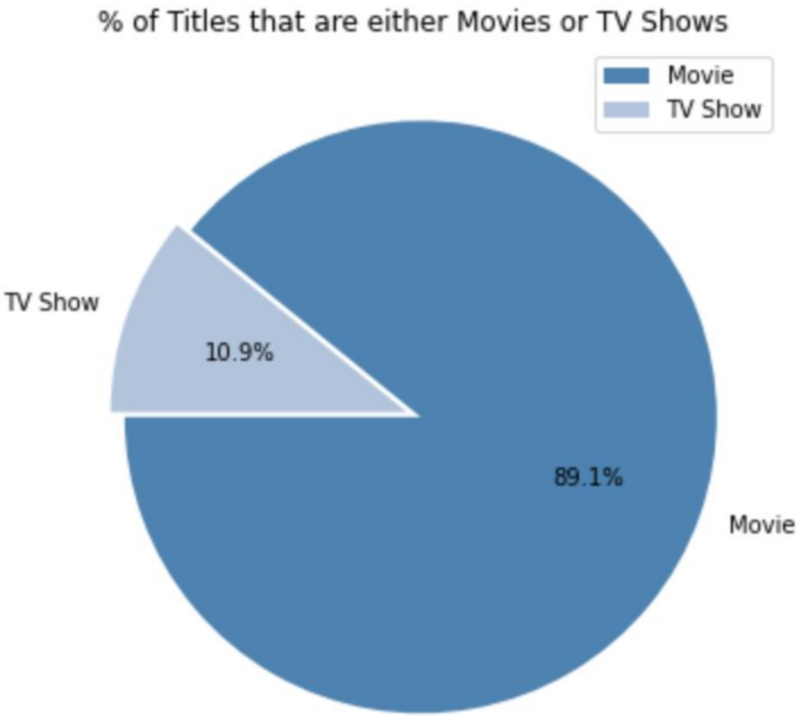
Recommendation System



Exploratory Data Analysis

I. General
II. Audience
III. Directors

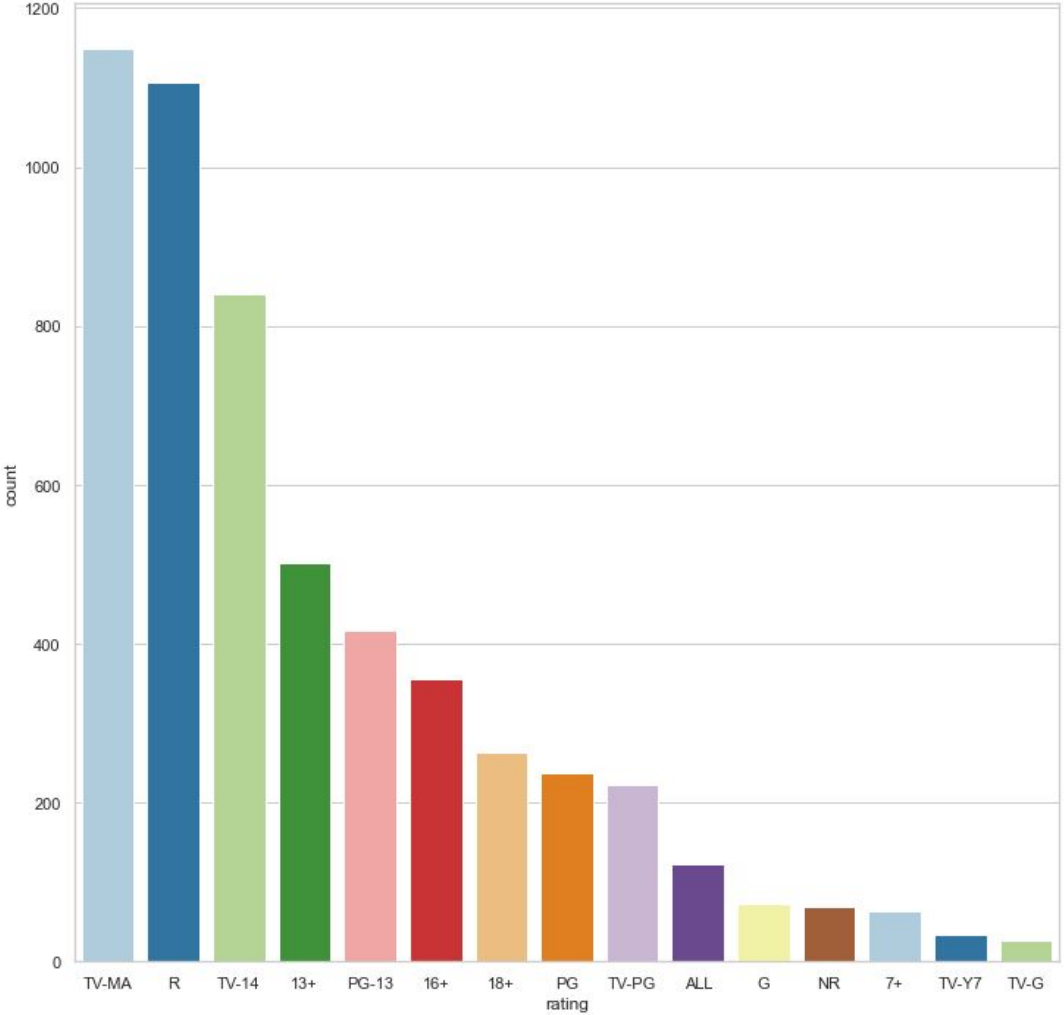
% of Titles that are Either Movies or TV Shows



Movie Ratings Analysis

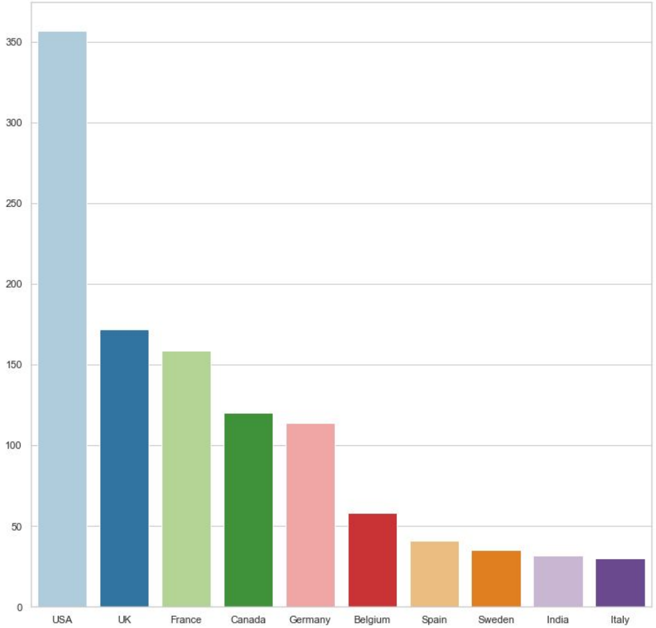
Which movie rating counts the most in our dataset

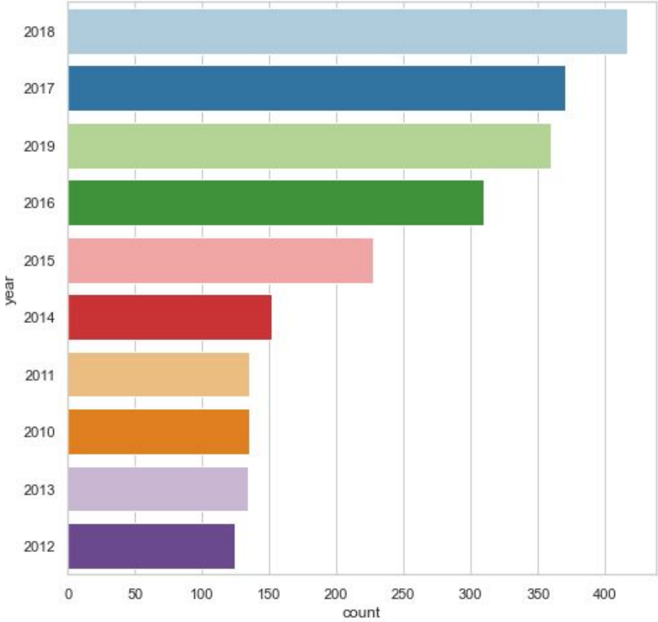
- 'TV-MA'
- 'R'
- 'TV-14'



COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

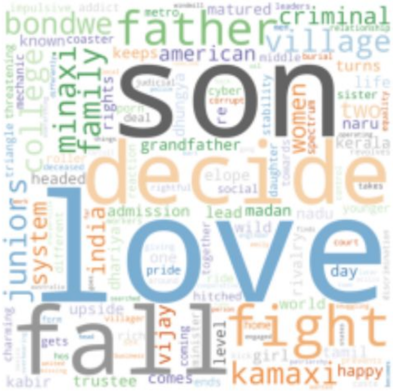
Top 10 Movie Content Creating Countries & Year Wise Analysis

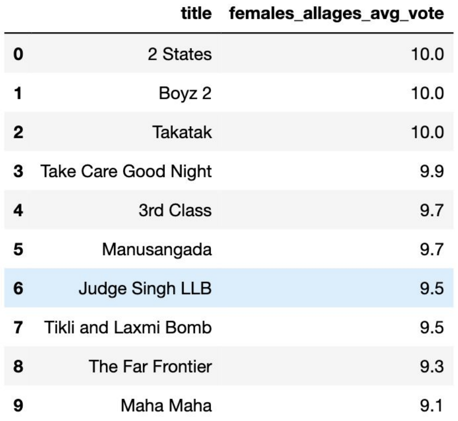




Word Cloud & Top 10 Rated Movies - By Gender

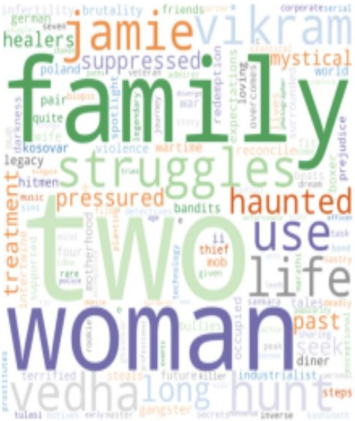
By Females

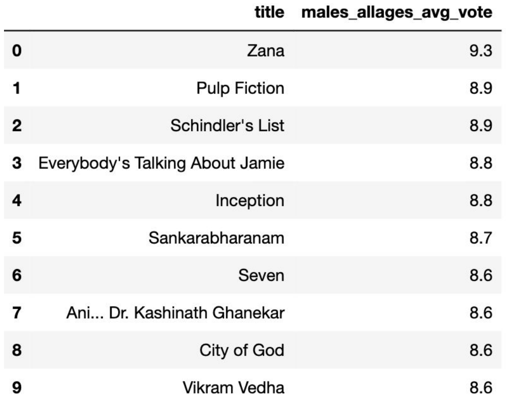




	title	females_allages_avg_vote
0	2 States	10.0
1	Boyz 2	10.0
2	Takatak	10.0
3	Take Care Good Night	9.9
4	3rd Class	9.7
5	Manusangada	9.7
6	Judge Singh LLB	9.5
7	Tikli and Laxmi Bomb	9.5
8	The Far Frontier	9.3
9	Maha Maha	9.1

By Males

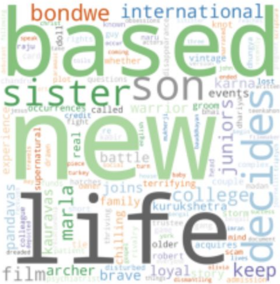


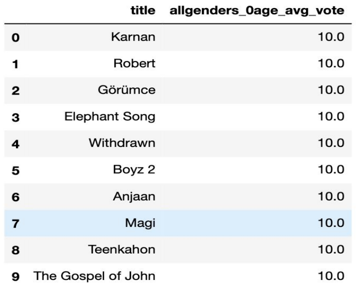


	title	males_allages_avg_vote
0	Zana	9.3
1	Pulp Fiction	8.9
2	Schindler's List	8.9
3	Everybody's Talking About Jamie	8.8
4	Inception	8.8
5	Sankarabharanam	8.7
6	Seven	8.6
7	Ani... Dr. Kashinath Ghanekar	8.6
8	City of God	8.6
9	Vikram Vedha	8.6

Word Cloud & Top 10 Rated Movies - By Age Group

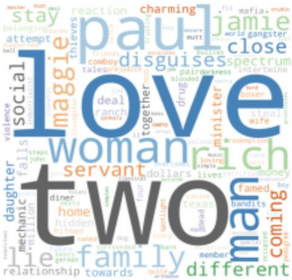
Below 18

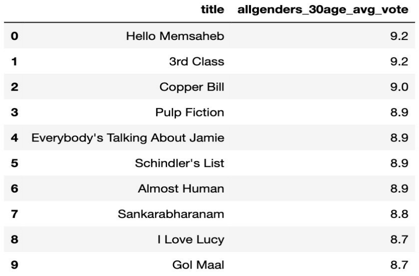




	title	allgenders_0age_avg_vote
0	Karnan	10.0
1	Robert	10.0
2	Görümce	10.0
3	Elephant Song	10.0
4	Withdrawn	10.0
5	Boyz 2	10.0
6	Anjaan	10.0
7	Magi	10.0
8	Teenkahon	10.0
9	The Gospel of John	10.0

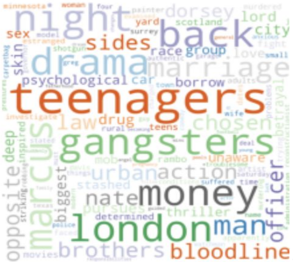
30 to 45

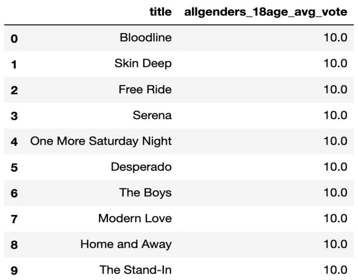




	title	allgenders_30age_avg_vote
0	Hello Memsaheb	9.2
1	3rd Class	9.2
2	Copper Bill	9.0
3	Pulp Fiction	8.9
4	Everybody's Talking About Jamie	8.9
5	Schindler's List	8.9
6	Almost Human	8.9
7	Sankarabharanam	8.8
8	I Love Lucy	8.7
9	Gol Maal	8.7

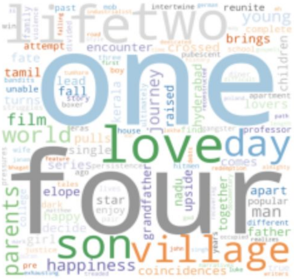
18 to 30

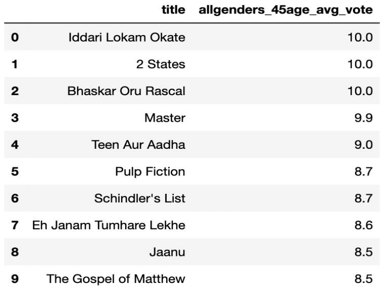




	title	allgenders_18age_avg_vote
0	Bloodline	10.0
1	Skin Deep	10.0
2	Free Ride	10.0
3	Serena	10.0
4	One More Saturday Night	10.0
5	Desperado	10.0
6	The Boys	10.0
7	Modern Love	10.0
8	Home and Away	10.0
9	The Stand-In	10.0

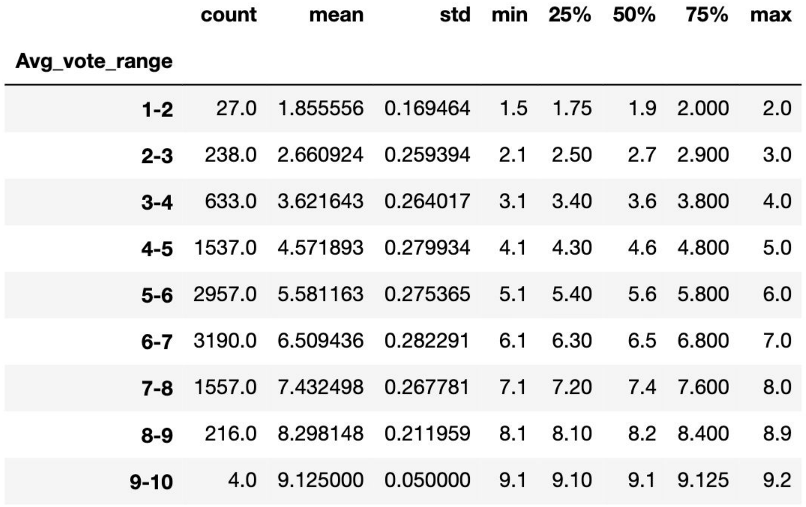
Above 45



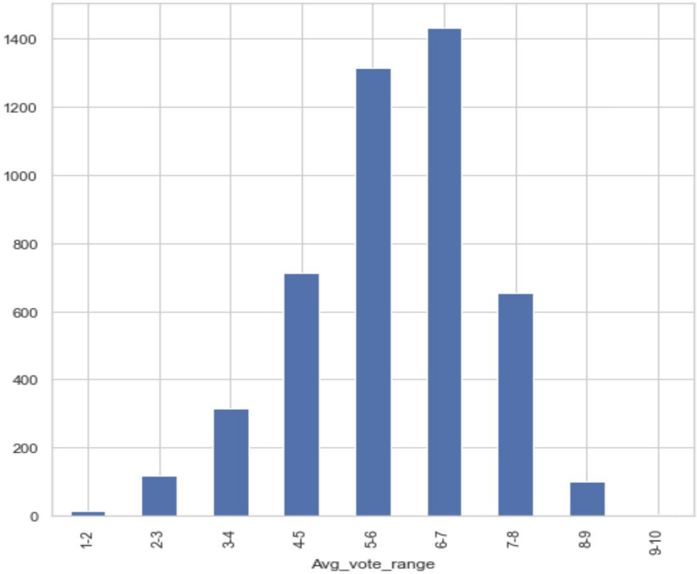


	title	allgenders_45age_avg_vote
0	Iddari Lokam Okate	10.0
1	2 States	10.0
2	Bhaskar Oru Rascal	10.0
3	Master	9.9
4	Teen Aur Aadha	9.0
5	Pulp Fiction	8.7
6	Schindler's List	8.7
7	Eh Janam Tumhare Lekhe	8.6
8	Jaanu	8.5
9	The Gospel of Matthew	8.5

Genre-Rating Analysis



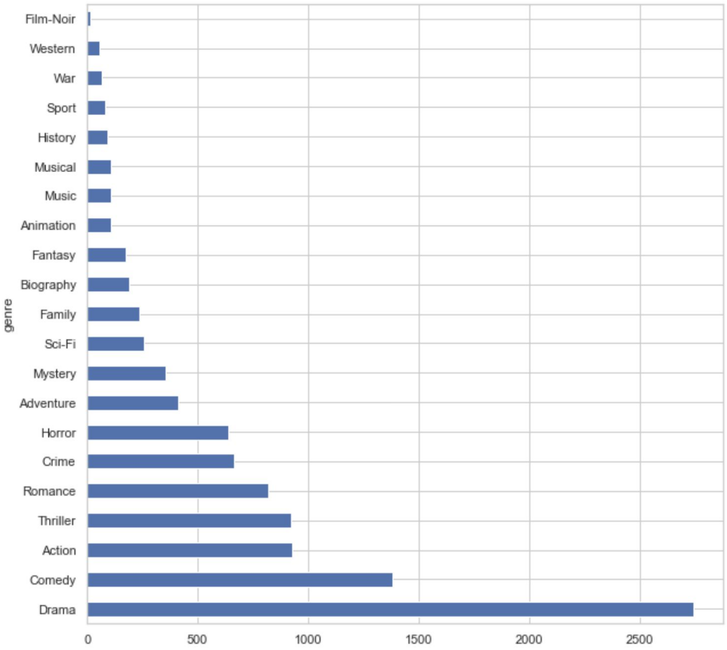
Avg_vote_range	count	mean	std	min	25%	50%	75%	max
1-2	27.0	1.855556	0.169464	1.5	1.75	1.9	2.000	2.0
2-3	238.0	2.660924	0.259394	2.1	2.50	2.7	2.900	3.0
3-4	633.0	3.621643	0.264017	3.1	3.40	3.6	3.800	4.0
4-5	1537.0	4.571893	0.279934	4.1	4.30	4.6	4.800	5.0
5-6	2957.0	5.581163	0.275365	5.1	5.40	5.6	5.800	6.0
6-7	3190.0	6.509436	0.282291	6.1	6.30	6.5	6.800	7.0
7-8	1557.0	7.432498	0.267781	7.1	7.20	7.4	7.600	8.0
8-9	216.0	8.298148	0.211959	8.1	8.10	8.2	8.400	8.9
9-10	4.0	9.125000	0.050000	9.1	9.10	9.1	9.125	9.2

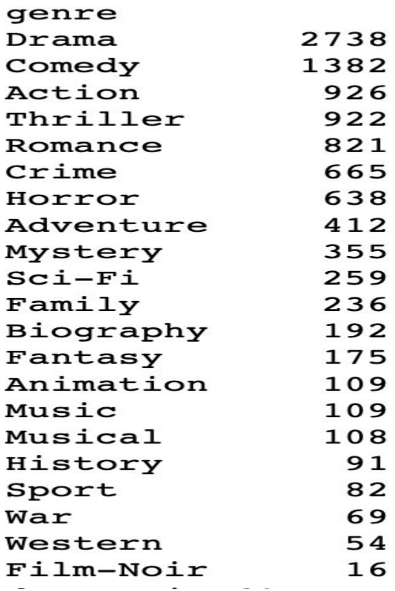


- **STEP 1:** Divide the average rating of movies from 1 to 10 into ten ranges

- **Median: 5-6   Mean: 6-7**

Genre-Rating Analysis

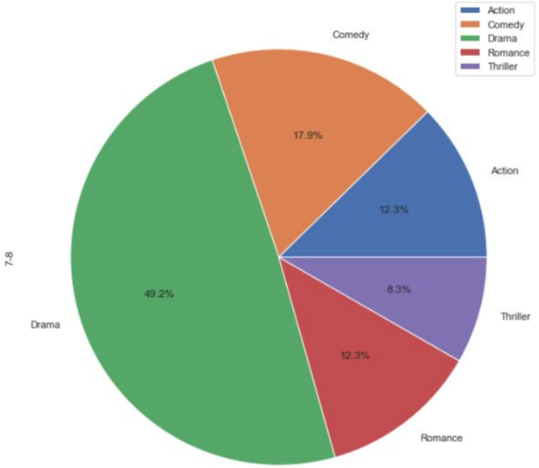


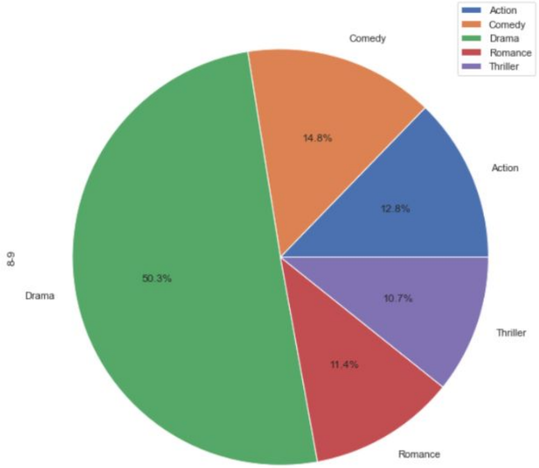


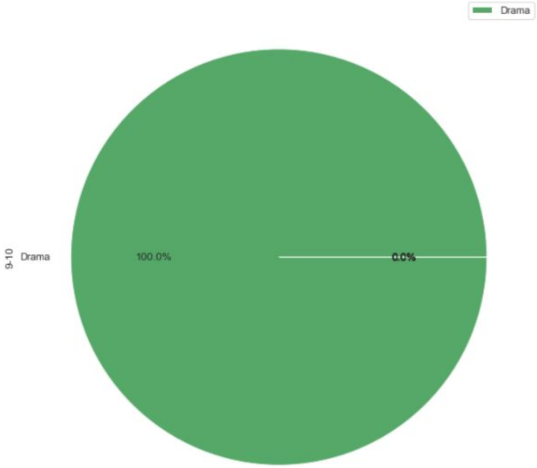
```
genre
Drama          2738
Comedy         1382
Action          926
Thriller        922
Romance         821
Crime           665
Horror          638
Adventure       412
Mystery         355
Sci-Fi          259
Family          236
Biography       192
Fantasy         175
Animation       109
Music           109
Musical         108
History          91
Sport            82
War              69
Western          54
Film-Noir        16
```

- **STEP 2:** Identify the number of genres and count every genre

Genre-Rating Analysis







- **STEP 3:** We only select movies with average votes higher than 7. Proportions are relatively stable

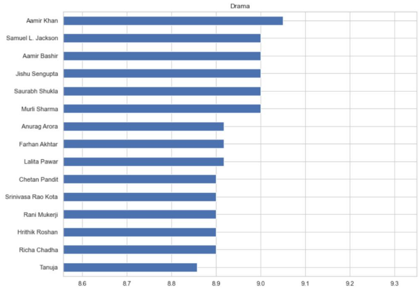
Genre-actors Recommendation Analysis

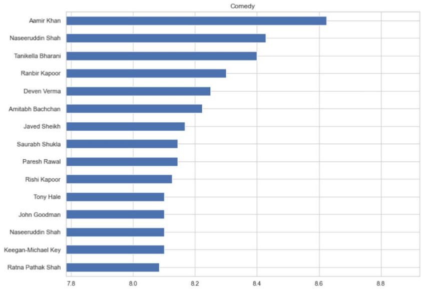
- Design for movie directors to use
  - When they selected a genre, they can view the corresponding top rated actors, so that they can choose among my top rated actors list
- Approach
  - count the size of each genre
    - Focus on the top 8 genres. They are namely:

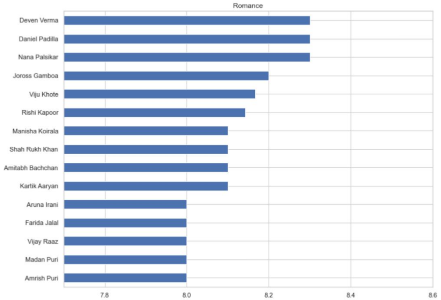
Drama	2738
Comedy	1382
Action	926
Triller	922
Romance	821
Crime	665
Horror	638
Adventure	412

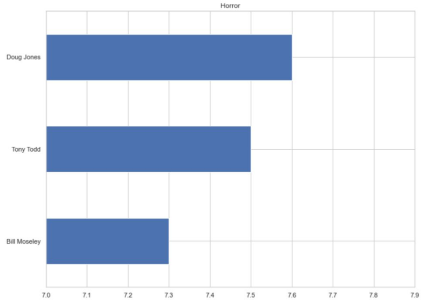
  - actors with 'avg_vote' greater than 9 are great actors for movie directors to consider
  - sort values by 'ave_rating_actor' in descending order and display top 15

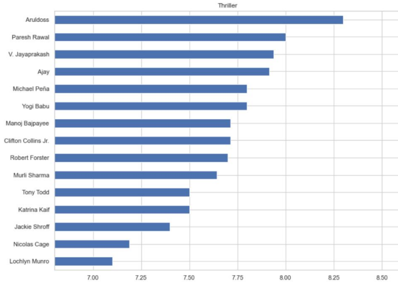
Genre-actors Recommendation Analysis

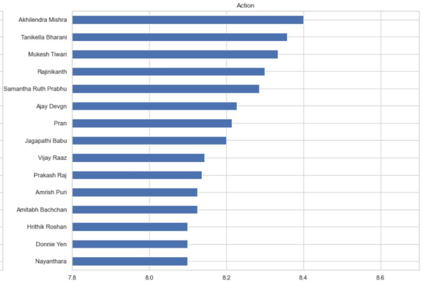


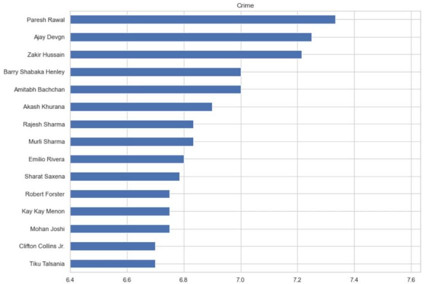


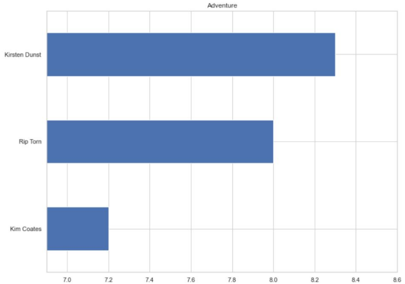




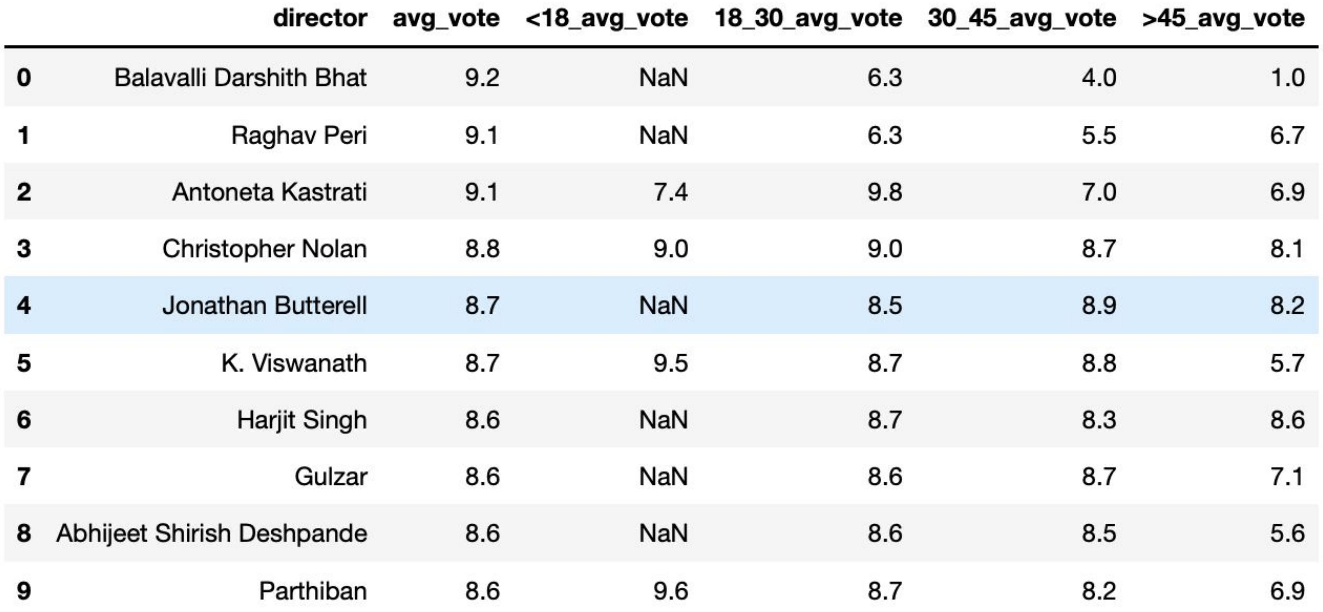








Top 9 Highest Rated Directors and Their Ratings by Age Groups



	director	avg_vote	<18_avg_vote	18_30_avg_vote	30_45_avg_vote	>45_avg_vote
0	Balavalli Darshith Bhat	9.2	NaN	6.3	4.0	1.0
1	Raghav Peri	9.1	NaN	6.3	5.5	6.7
2	Antoneta Kastrati	9.1	7.4	9.8	7.0	6.9
3	Christopher Nolan	8.8	9.0	9.0	8.7	8.1
4	Jonathan Butterell	8.7	NaN	8.5	8.9	8.2
5	K. Viswanath	8.7	9.5	8.7	8.8	5.7
6	Harjit Singh	8.6	NaN	8.7	8.3	8.6
7	Gulzar	8.6	NaN	8.6	8.7	7.1
8	Abhijeet Shirish Deshpande	8.6	NaN	8.6	8.5	5.6
9	Parthiban	8.6	9.6	8.7	8.2	6.9

Our big concept

Build a movie recommendation machine for movie lovers and a popularity prediction system for movie makers

# 1 Movie Recommender

Building the Movie Recommender

1 Review and Preprocess the Dataset

5731 Movies with description, cast, genre, writer and director

2 Convert Text Columns into Numerical Values

Use TfIdfVectorizer from scikit-learn to construct TF-IDF matrix
Use CountVectorizer from sklearn

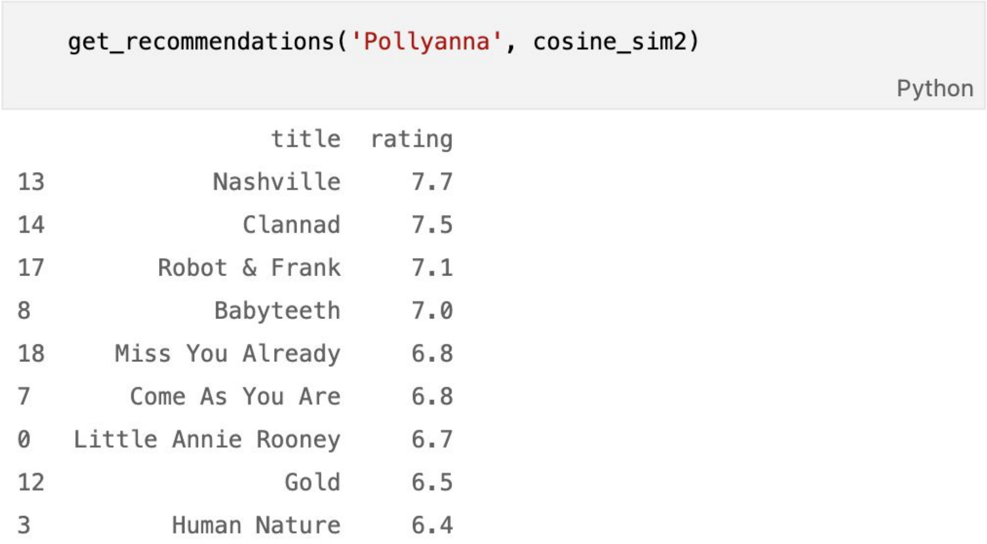
3 Measure Similarity between Movies

Cosine Similarity measures the cosine of the angle between two vectors

4 Define Recommendation Function

Get pairwise similarity scores of all movies with that movie input;
Return 10 most popular movies out of 20 most similar movies

An Example of the Movie Recommender



```
get_recommendations('Pollyanna', cosine_sim2)
```

Python

```
               title  rating
13         Nashville     7.7
14           Clannad     7.5
17     Robot & Frank     7.1
8          Babyteeth     7.0
18  Miss You Already     6.8
7       Come As You Are     6.8
0  Little Annie Rooney     6.7
12              Gold     6.5
3       Human Nature     6.4
```

Inputs

A Movie You Enjoy

Cosine Similarity Matrix

Outputs

Top 10 Recommended Movies

- Similarity
- Popularity

# 2 Popularity Prediction System

Building Popularity Prediction System for Movie Maker: Can you beat the average?

1 Data Preparation

Use only data for movies
Get dummies for duration range and genre

2 Web scraping for directors, actors and production company

Scrap from THE NUMBERS website
for the ranking of worldwide box office

3 Data Processing

Segment directors, actors, and production companies by quantiles of ranking into dummies;
Define movies rated higher than average as "high vote" that beat the average market rating

4 Define Recommendation Function

Given the duration, genre, director, actor, and production company that a movie maker inputs
-> **Classification: whether the movie can beat the average**

Machine Learning Model for Popularity Prediction System

Independent variables

Explanatory Variables(All Dummies)	Baseline to avoid perfect collinearity
DURATION RANGE: ‘< 95 min’, ‘> 120 min’	‘95 - 120 min’
GENRE: 'Adventure', 'Animation', 'Biography', 'Comedy', 'Crime', 'Drama', 'Family', 'Fantasy', 'History', 'Horror', 'Music', 'Musical', 'Mystery', 'Romance', 'Sci-Fi', 'Sport', 'Thriller', 'War', 'Western'	‘Action’
DIRECTOR RANK RANGE: '25th_to_50th_percentile_director','50th_to_75th_percentile_director','75th_to_100th_percentile_director'	‘0th_to_25th_percentile_director’
ACTOR RANK RANGE: '25th_to_50th_percentile_actor','50th_to_75th_percentile_actor','75th_to_100th_percentile_actor'	‘0th_to_25th_percentile_actor’
PRODUCTION COMPANY RANK RANGE: '25th_to_50th_percentile_production','50th_to_75th_percentile_production','75th_to_100th_percentile_production'	‘0th_to_25th_percentile_production’

Dependent variables

0 – “Sorry, your movie will not reach the average movie rating in the market :(Try another combination!”

1 – “Congratulations, your movie will beat the average rating in the market !”

Train: Test = 7 : 3

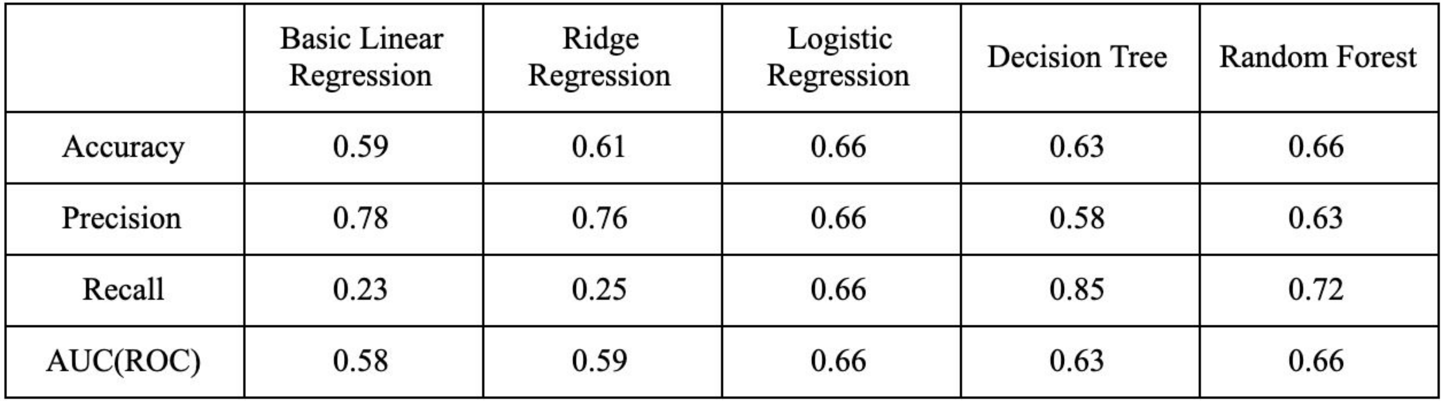
Results

Determination rule for threshold(0.7):

Hold a level of precision score(0.65) and maximize others

Regression models

Classification models

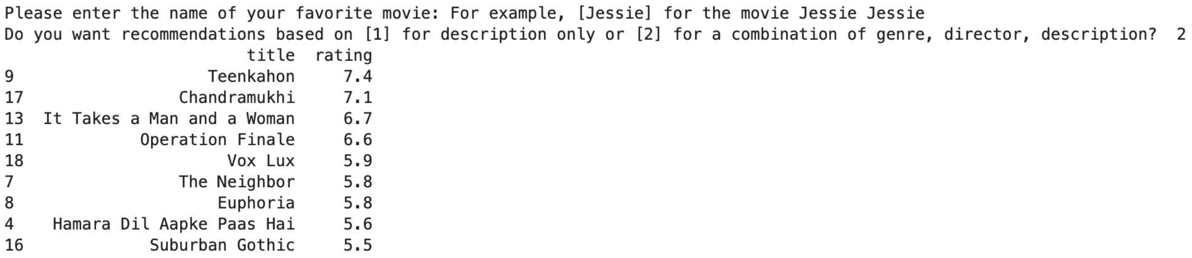


	Basic Linear Regression	Ridge Regression	Logistic Regression	Decision Tree	Random Forest
Accuracy	0.59	0.61	0.66	0.63	0.66
Precision	0.78	0.76	0.66	0.58	0.63
Recall	0.23	0.25	0.66	0.85	0.72
AUC(ROC)	0.58	0.59	0.66	0.63	0.66

What's more: Cross Validation & Feature Importance to better analyze the data and understand our model

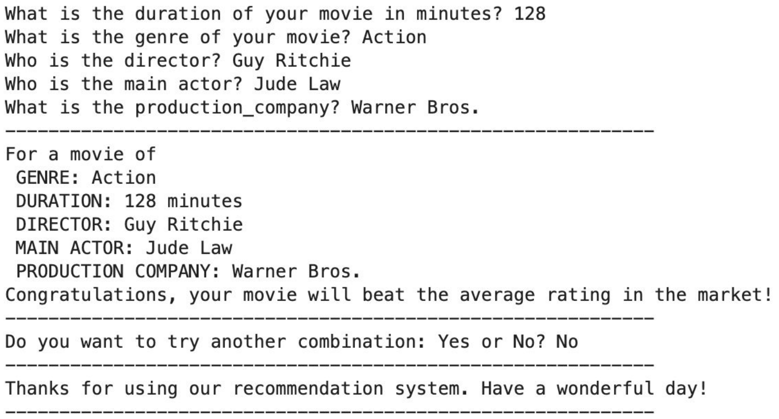
Brief Overview of the Machine

For movie lovers



```
Please enter the name of your favorite movie: For example, [Jessie] for the movie Jessie Jessie
Do you want recommendations based on [1] for description only or [2] for a combination of genre, director, description?  2
                         title  rating
9                    Teenkahon     7.4
17                 Chandramukhi    7.1
13   It Takes a Man and a Woman    6.7
11              Operation Finale   6.6
18                      Vox Lux    5.9
7                  The Neighbor    5.8
8                      Euphoria    5.8
4      Hamara Dil Aapke Paas Hai   5.6
16              Suburban Gothic    5.5
```

For movie makers



```
What is the duration of your movie in minutes? 128
What is the genre of your movie? Action
Who is the director? Guy Ritchie
Who is the main actor? Jude Law
What is the production_company? Warner Bros.
------------------------------------------------------------
For a movie of
 GENRE: Action
 DURATION: 128 minutes
 DIRECTOR: Guy Ritchie
 MAIN ACTOR: Jude Law
 PRODUCTION COMPANY: Warner Bros.
Congratulations, your movie will beat the average rating in the market!
------------------------------------------------------------
Do you want to try another combination: Yes or No? No
------------------------------------------------------------
Thanks for using our recommendation system. Have a wonderful day!
------------------------------------------------------------
```

Thanks!

Any **questions** ?